修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　電気通信　学研究科　システム工学　専攻　博士前期課程		
氏　　名	仁科　朋也	学籍番号	0835032
論 文 題 目	検索支援のための Web 文書クラスタリング手法		

要　　旨

今日，Web からユーザの望む情報を得る手段としてサーチエンジンが利用される．しかし、ユーザが望む情報を持たないページも多数表示されるため，各ページがユーザの望む情報を含むかどうかを判断するのに時間と労力を割かなければならない．このような負担を軽減するための検索支援手法として，検索結果を分類して表示する Web 文書クラスタリングが挙げられる．

Web 文書クラスタリング手法として，抽出された各重要語を含む Web ページ集合をひとつの文書クラスタとする手法が広く用いられている．しかし，従来の研究では重要語間の類似度を考慮していないために，類似した話題を表す語句が重要語として抽出されると，話題が類似するクラスタが複数出力されてしまうという欠点がある．そこで本研究ではこの問題点を解消するために，単語間の類似度を考慮した Web 文書クラスタリング手法を提案する．本手法は，サーチエンジンが返すタイトルとスニペットの単語分布情報から，互いに類似していない重要語を抽出する．続いてどのクラスタにも属さない Web ページをできるだけ減らすために，重要語から直接文書クラスタを生成せずに，各重要語に類似した単語集合を単語クラスタとして生成し，単語クラスタから文書クラスタを生成する．次に生成した文書クラスタの内容を把握しやすくするために，クラスタ名と説明文を生成する．クラスタ名生成手法として節と節の係り受けを利用する手法があるが，サーチエンジンの検索結果は情報量が少ないのでこの手法は向いていない．そこで本手法ではクラスタ名は複合名詞や名詞句といった短い表現でクラスタの内容を表す．また本手法の説明文は重要な単語を説明している文とする．新聞記事を対象として「～は～である」など特定の表現を利用する既存手法があるが，情報量が少ない，Web ページには様々な形式の表現が存在するなどの理由でこの手法を用いることは難しい．そこで「～は」などを区切り文字として文節に区切り，区切り文字に重みをつけることにより説明文を生成する．

そして，従来手法（語句間の類似度を考慮しない方法）との比較評価およびユーザによる評価を行う．比較評価では実際に人手で分類したものを正解データとし，本手法のほうがクラスタリングの精度，被覆度ともに優れていること，生成されるクラスタのサイズや重複割合も適切であることを示す．ユーザによる評価では本手法を実装したシステムをユーザに提示し，クラスタ名，説明文によりクラスタの内容が把握しやすくなること，普段使用しているサーチエンジンと比べて本手法を実装したシステムの方が有用であることを示す．